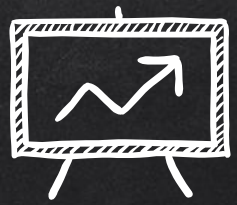


Basic Statistic

--Concepts/Knowledge/Practice

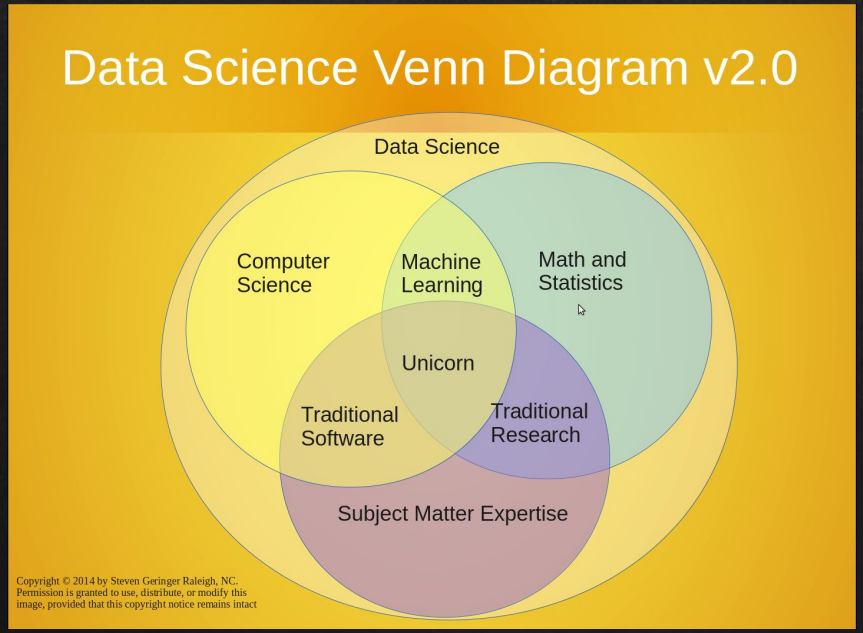
主讲人：Lucy Xu

2021

主讲人介绍

- 10+年数据科学，预测建模等领域工作经验，在消费者贷款领域拥有丰富的经验及知识，包括房贷， 车贷，信用卡和个人贷款。
- 现任全美前五保险公司，机器学习和人工智能部门 – Data Science Director，专注于IoT(internet of things，物联网) for Connected car/home insurance product 和 NLP(Natural Language Processing)
- 曾任职CapitalOne数据科学经理和DCF Global – Senior Data Science Manager. 带领团队使用大数据和机器学习，为整个信贷业务领域及产业链提供服务及解决方案。专注于市场营销，信用风险，定价，估值，运营，帐户管理和欺诈检测。
- 本科毕业于中国科学技术大学，在美国分别获得生物物理的博士和统计硕士学位

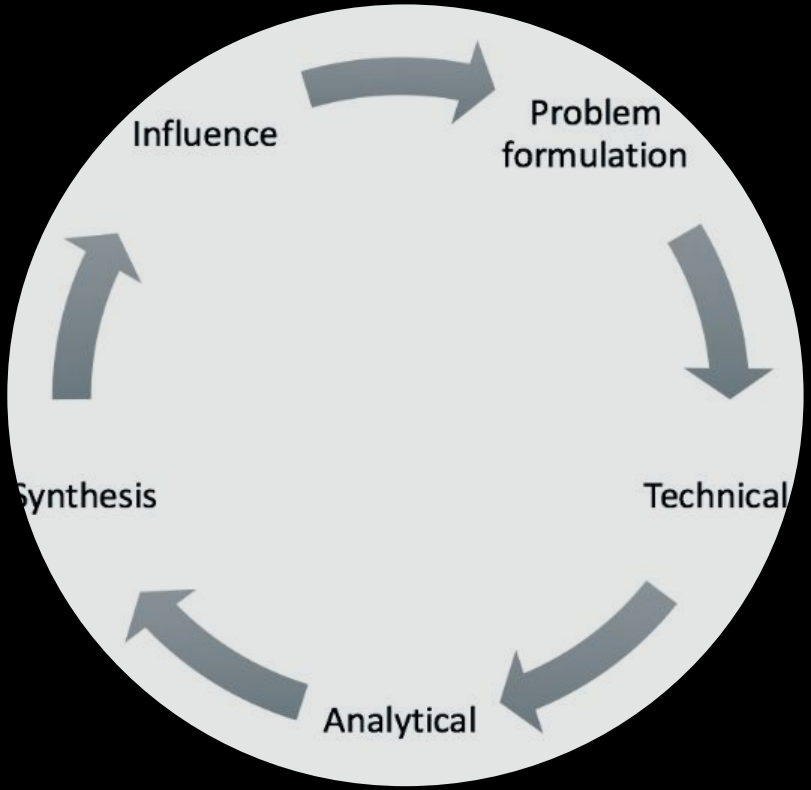
Statistics and Math are Heavily Weighted for Data Scientist Skill Sets



"Data Scientist is a person who is better at statistics than any programmer and better at programming than any statistician."

Battle of Data Science definition

Five Core Skills Data Scientist Manager is Looking for



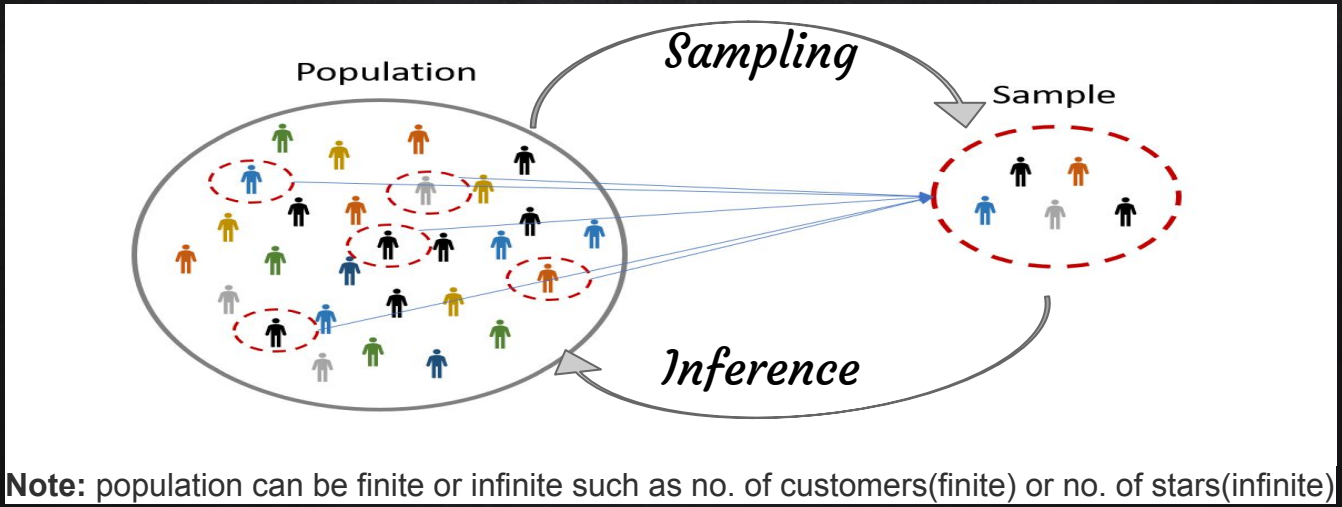
Statistics and Probability are Building Blocks of All Data Heavy Practice

"Statistics is a mathematical science pertaining to the collection, analysis, interpretation or explanation, and presentation of data"

Start with Some Basic Terminologies – 1

Population: "The entire set of individuals or objects of interest"

Sample: "A portion or a part of the population of interest."



Note: population can be finite or infinite such as no. of customers(finite) or no. of stars(infinite)

Start with Some Basic Terminologies – 2

Variable: "A characteristic of each individual element of a population/sample".

Data: "A value of the variable associated with one element of a population or sample.it can be numbers, words or symbols."

Experiment: "A planned activity whose result yields a set of data."

Parameter: "A numerical value which summarizes the entire population or the descriptive measure of the population. For example, population mean, population variance, population standard deviation, etc."

Statistic: "A numerical value which summarizes the sample data or descriptive measure of the sample. For example, sample mean, sample variance, sample standard deviation, etc. (Descriptive statistics and Inference statistics)"

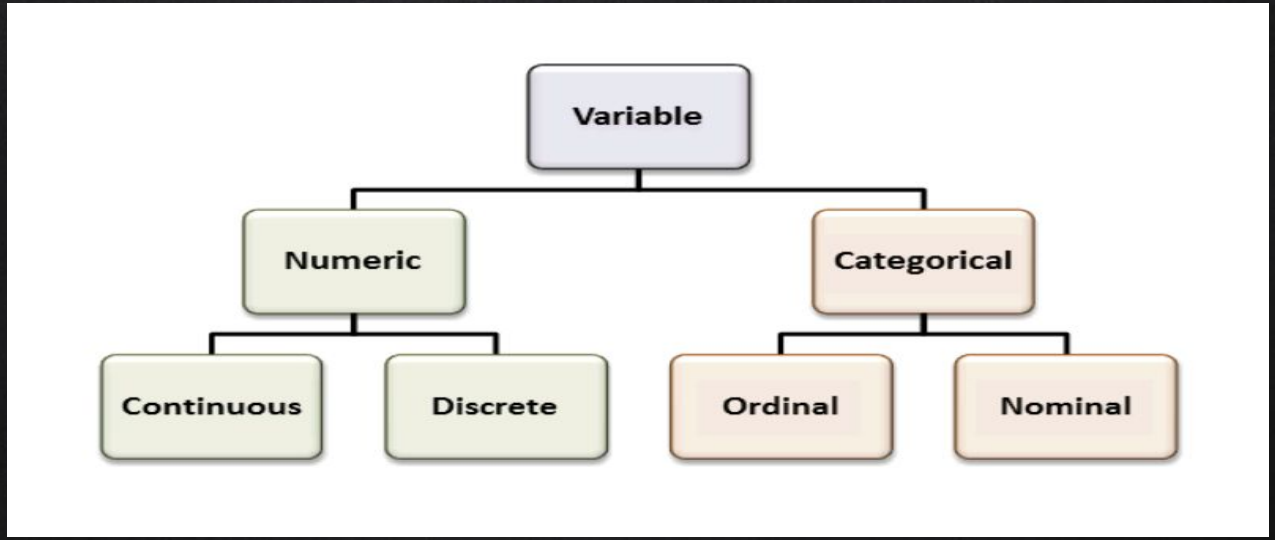
Table of Contents for Today's Discussion

1. Basics Statistical concepts/methods to understand data (EDA statistics)

2. Sampling Techniques

3. Probability distribution and Inferential statistical analysis

Variable/Data Types



Independent(Features) variable and dependent(Target/Outcome) variable

Practice

Please label the data type, quantity/quality and continuous, for the list of variables (2-3 mins practice)

Sex,

State,

Grades(A/B/C/D),

number of children in a household,

monthly income,

Product rating (not recommend, neutral, may recommend, strongly recommend)

Company expenses

Customer ID

Practice – Answer

Please label the data type, quantity/quality and continuous, for the list of variables (2–3 mins practice)

Sex – qualitative, Nominal

State – qualitative, Nominal

Grades(A/B/C/D) – qualitative, ordinal

number of children in a household – quantitative, discrete

monthly income – quantitative, continuous

Product rating (not recommend, neutral, may recommend, strongly recommend) – qualitative, ordinal

Company expenses – quantitative, continuous

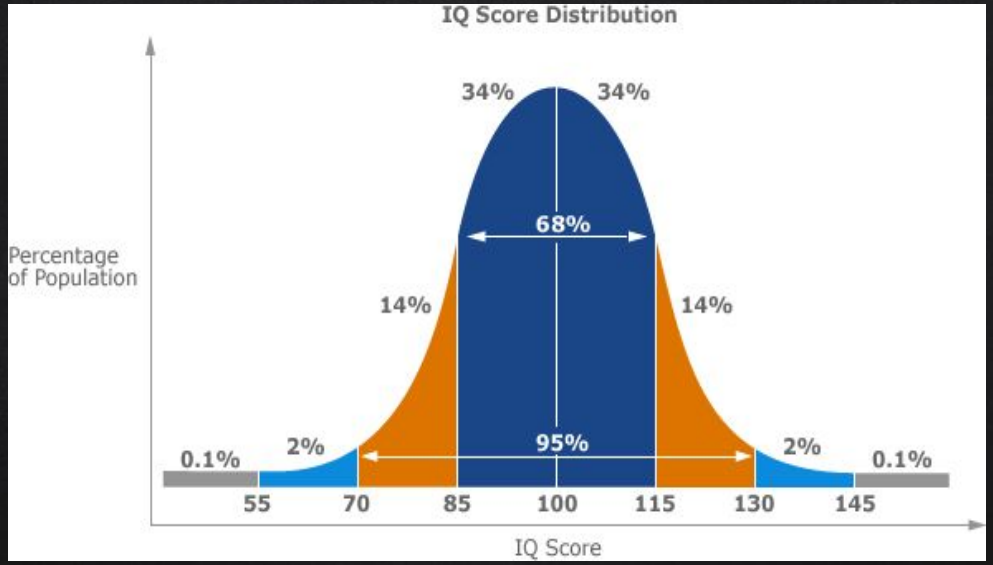
Customer ID – qualitative, Nominal (caution in real life application)

Descriptive Statistics – 1

- Measure of central tendency
  - Mean: Measure of the average of all the values in a sample is called Mean $$\text{Mean} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$
  - Median: Measure of the central value of the sample set is called Median. (Is ρ50 robust to data skewness, is a robust estimator)
  - Mode: The value most recurrent in the sample set is known as Mode. (Used frequently for categorical variable, sometime used for discrete numerical)
  - Weighted Mean (Two major reasons to use it) $$\text{Weighted mean} = \bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- Measure of spreadness
  - *Variance*
  - *Stand deviation*
  $$\text{Variance} = s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$
  $$\text{Standard deviation} = s = \sqrt{\text{Variance}}$$
  - Quartile: Quartiles tell us about the spread of a data set by breaking the data set into quarters, just like the median breaks it in half (Robust).
  - Range: It is the given measure of how spread apart the values in a data set are. The range can be calculated as: *Range = Max(x_i) – Min(x_i)*

Descriptive Statistics – 1 Example



Descriptive Statistics – 1 Practice

Practice: The salaries of staff members is given below:

Case 1: Staff salary = 15k, 18k, 16k, 14k, 15k, 15k, 12k, 17k, 90k, 95k

Mean? Median? Mode? Percentiles?

Mean	Median	Mode	Min	P25	P50	P75	Max

Case 2: Staff salary = 15k, 18k, 16k, 14k, 13k, 15k, 17k, 17k, 14k, 14k (bold one's are outliers)

Mean? Median? Mode? Percentile

Mean	Median	Mode	Min	P25	P50	P75	Max

$$\text{Mean} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Descriptive Statistics – 1 Practice Answer

Practice: The salaries of staff members is given below:

Case 1: Staff salary = 15k, 18k, 16k, 14k, 15k, 15k, 12k, 17k, 90k, 95k

Mean? Median? Mode? Percentiles?

Mean	Median	Mode	Min	P25	P50	P75	Max
30.7K	15.5K	15K	12K	15K	15.5K	18K	95K

Case 2: Staff salary = 15k, 18k, 16k, 14k, 13k, 15k, 17k, 17k, 14k, 14k (bold one's are outliers)

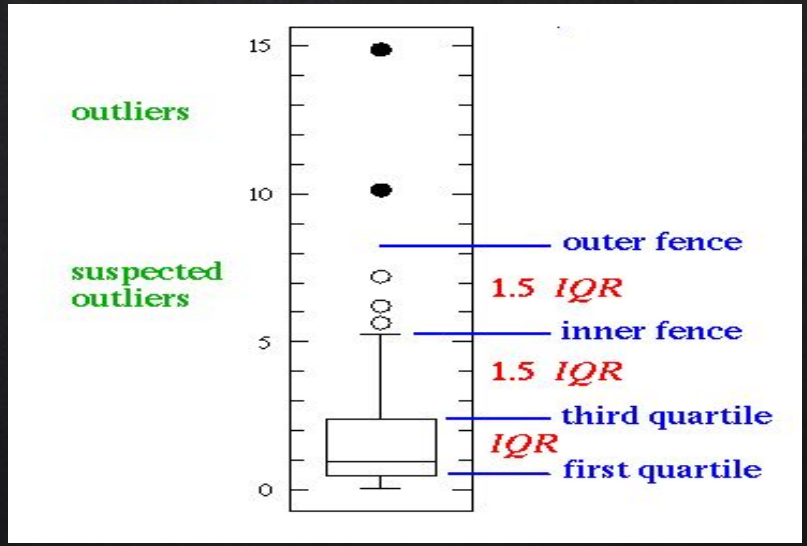
Mean? Median? Mode? Percentile

Mean	Median	Mode	Min	P25	P50	P75	Max
15.3K	15K	14K	13K	14K	15K	17K	18K

$$\text{Mean} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

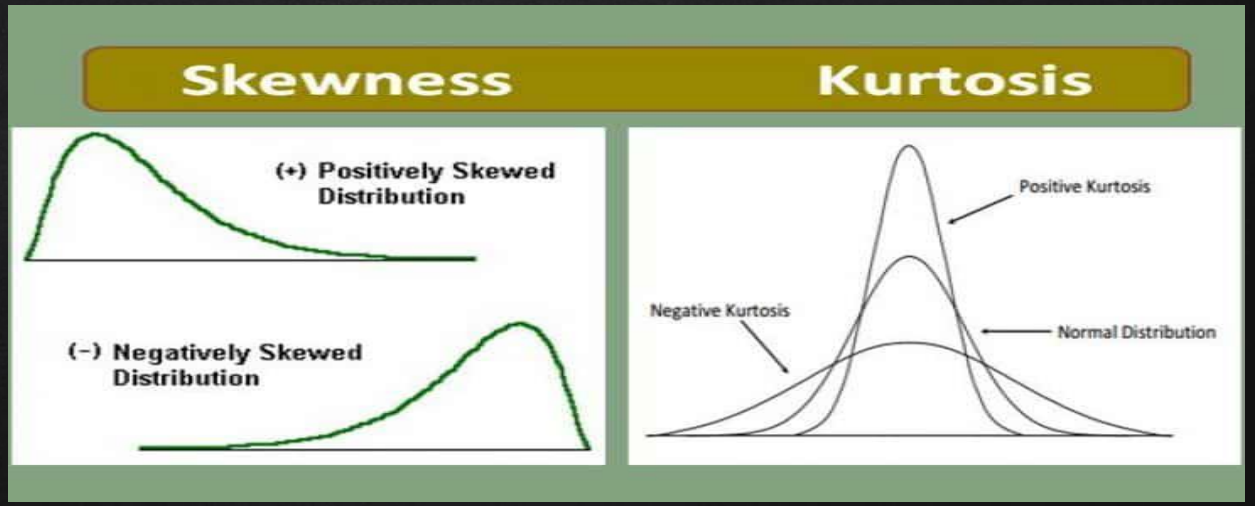
Descriptive Statistics – 2

- Outlier and IQR(Interquartile Range)
- Five number summary statistics – Boxplot (One of the simplest data distribution)



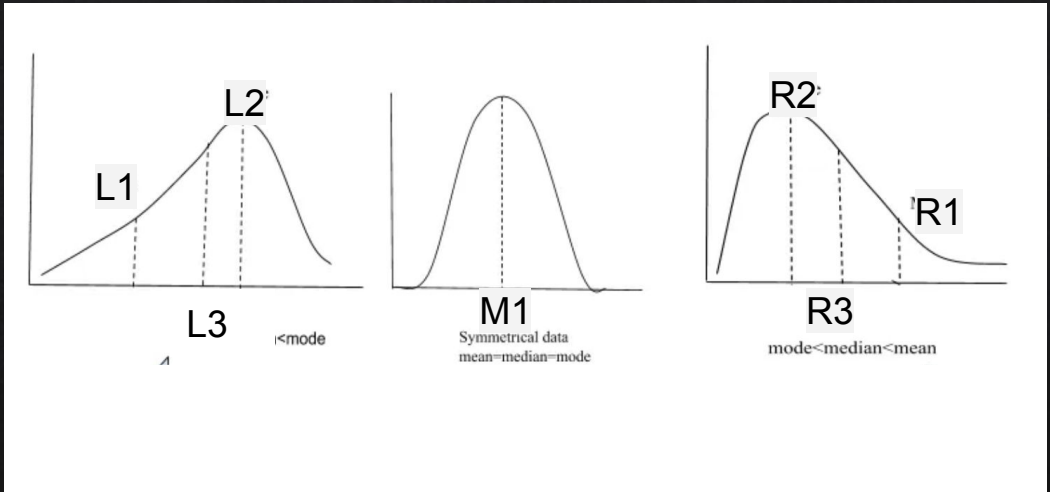
Descriptive Statistics for Exploring Data Distribution

- Skewness
- Kurtosis

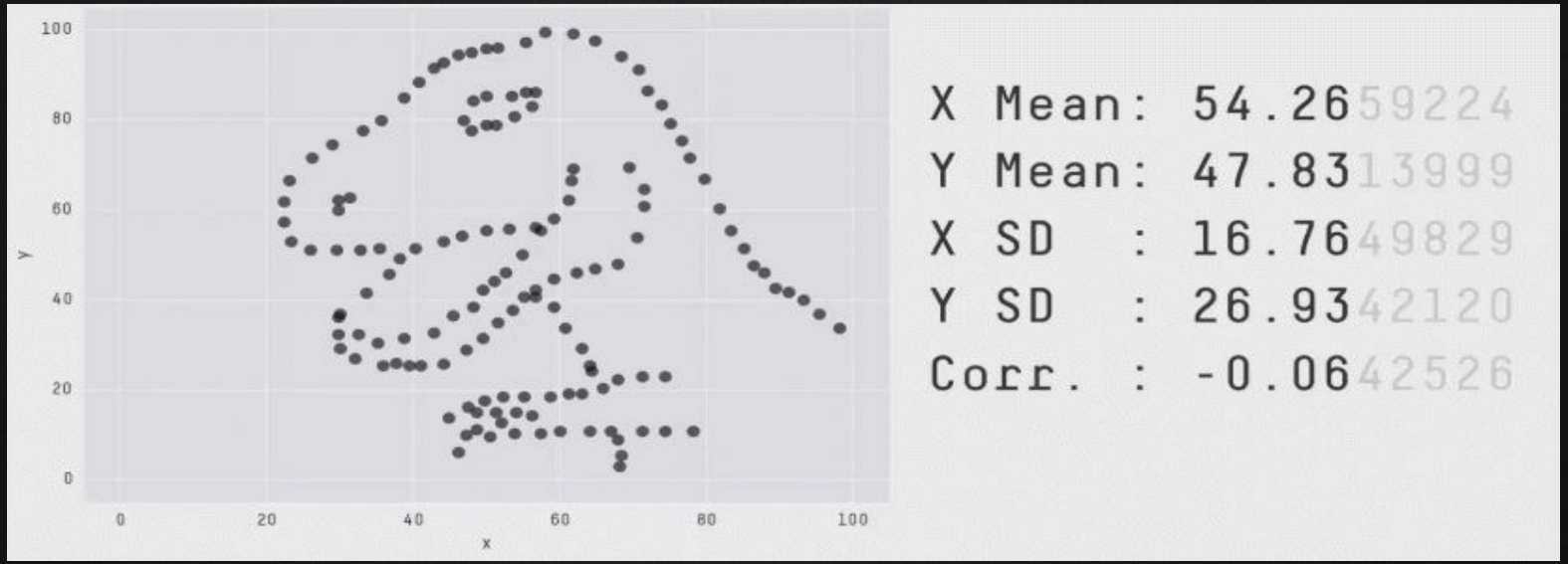


Exploring Data Distribution – Practice

Could you label Mean, Median and Mode on the three charts ?

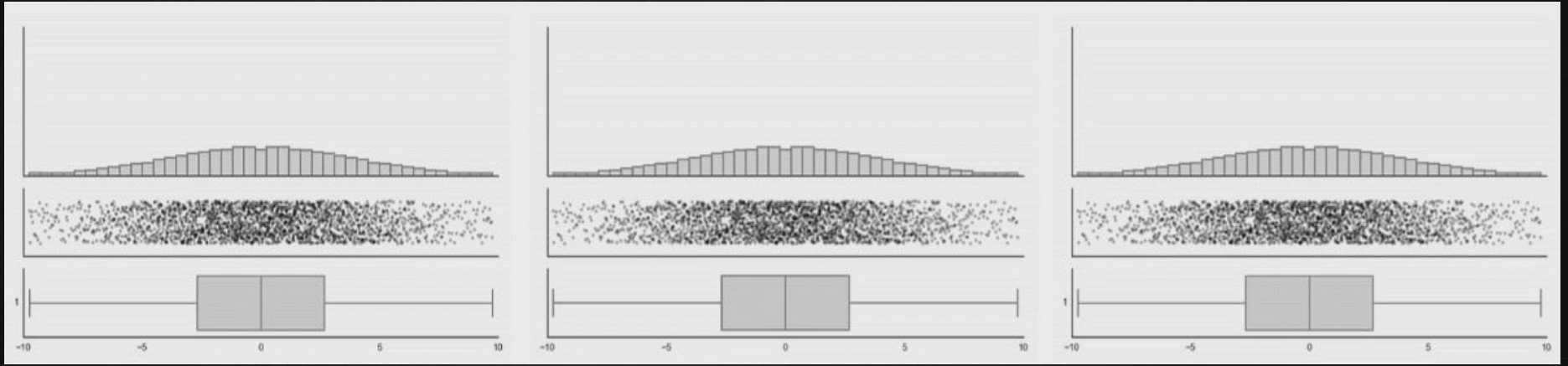


Is Mean and Variance Sufficient to Summarize/Understand your data?



https://www.autodeskresearch.com/publications/samestats

Is Boxplot Sufficient?



https://www.autodeskresearch.com/publications/samestats

Key Charts to Explore Data Distribution – Univariate

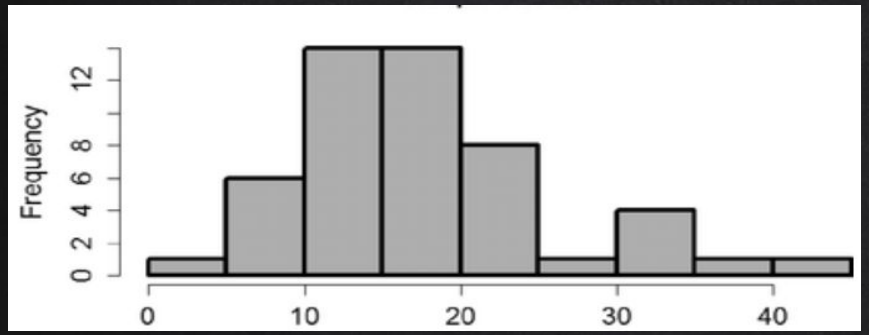
Boxplot
"A plot introduced by Tukey as a quick way to visualize the distribution of data"

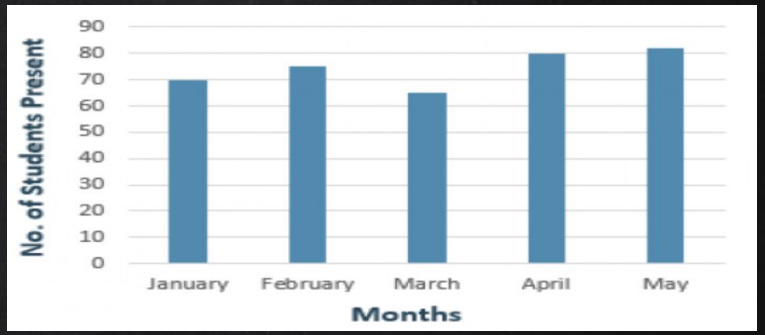
Frequency table/Bar Chart
"A tally of the count of categorical data for each category (numerical variable is histogram)"

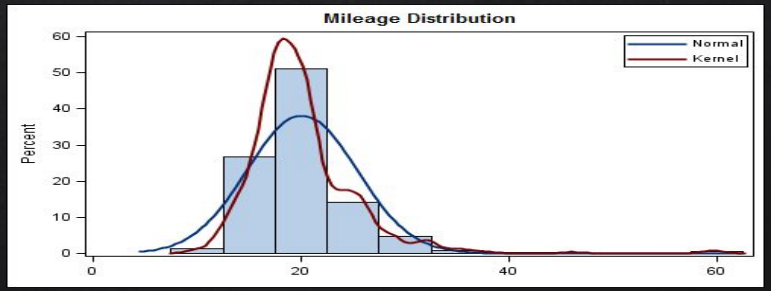
Histogram
"A plot of the frequency table with the bins on the x-axis and the count (or proportion) on the y-axis. While visually similar, bar charts should not be confused with histograms"

Density plot
"A smoothed version of the histogram, often based on a *kernel density estimate*"

How These Charts Looks Like







Key Charts to Explore Data Distribution – Bivariate

Scatter plot - correlation (two numerical variables)
"A plot in which the x-axis is the value of one variable, and the y-axis the value of another"

Heatmap - correlation (two numerical variables)
"A plot build on correlation matrix, showed the any two variable correlation in a heatmap color theme"

Correlation matrix
"A table where the variables are shown on both rows and columns, and the cell values are the correlations between the variables"

Contour plot
"A plot showing the density of two numeric variables like a topographical map"

Violin plot
"Similar to a boxplot but showing the density estimate"

Bar plot
"A boxplot for several bars showed one numerical measure across a nominal variable category"

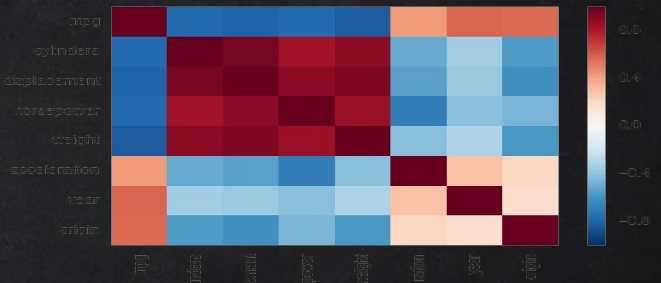
Contingency table
"A tally of counts between two or more categorical variables"

Bivariate Charts Example1

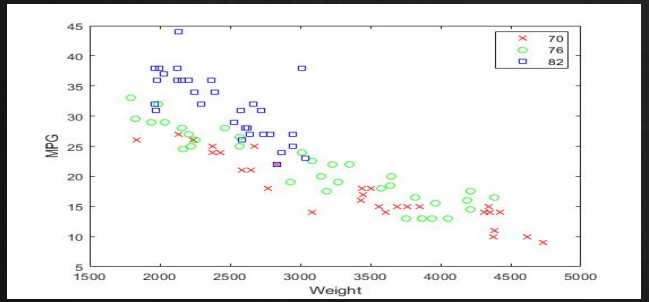
Correlation Matrix

	T	CTL	FTR	VZ	LVLT
T	1.000	0.475	0.328	0.678	0.279
CTL	0.475	1.000	0.420	0.417	0.287
FTR	0.328	0.420	1.000	0.287	0.260
VZ	0.678	0.417	0.287	1.000	0.242
LVLT	0.279	0.287	0.260	0.242	1.000

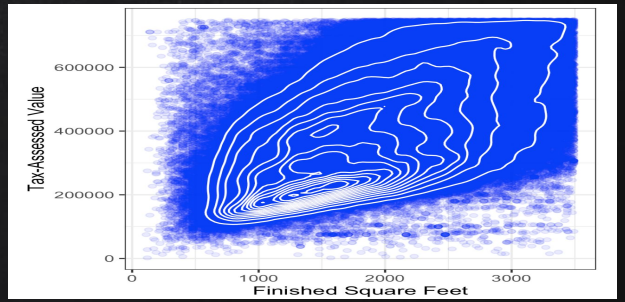
Correlation Heatmap



Scatter Plot



Contour Plot

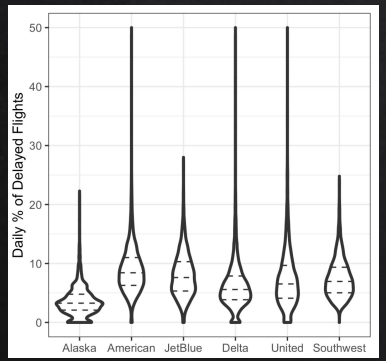


Bivariate Charts Example2

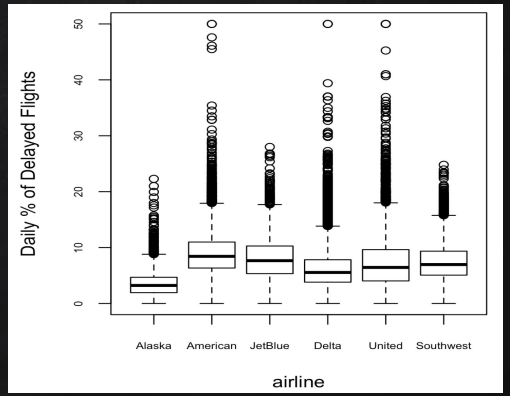
Contingency Table

	Dog	Cat	Total
Male	42	10	52
Female	9	39	48
Total	51	49	100

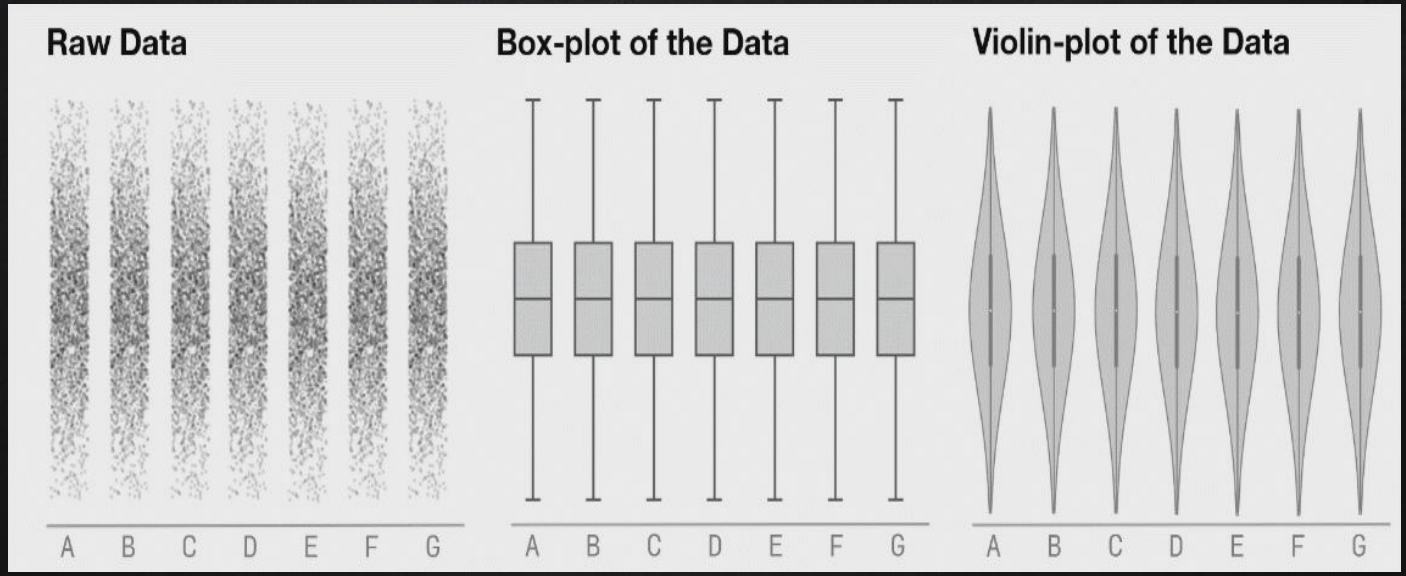
Violin Plot



Bar Plot



Bivariate Boxplot Normally OK ...



https://www.autodeskresearch.com/publications/samestats

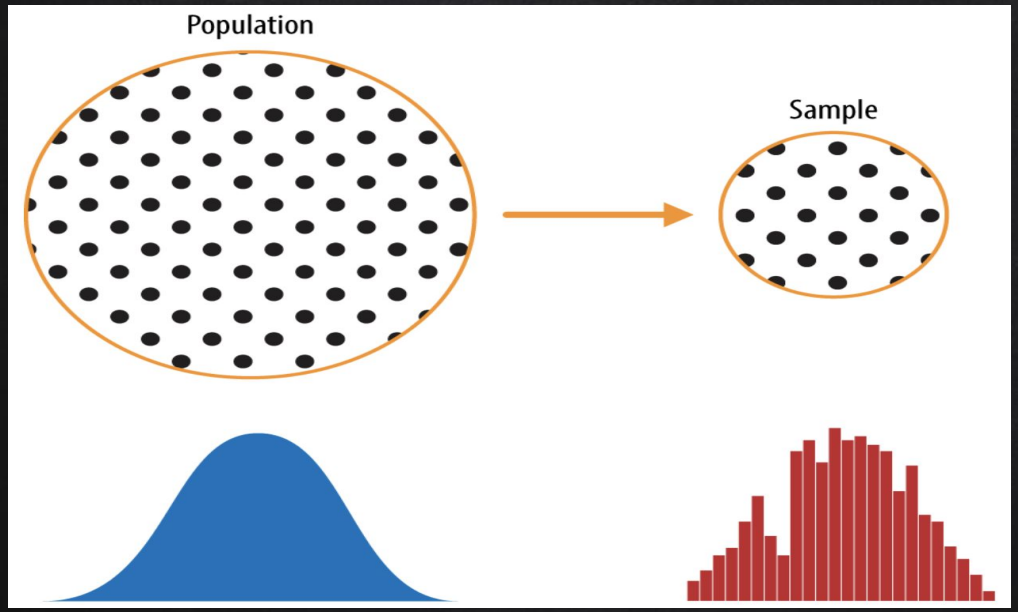
Table of Contents for Today's Discussion

1. Basics Statistics to understand data (EDA statistics)

2. Sampling Techniques

3. Probability distribution and Inferential statistical analysis

Population/Sample distribution



"Data quality often matters more than data quantity when making an estimate or a model based on a sample. Data quality in data science involves completeness, consistency of format, cleanliness, and accuracy of individual data points. Which means *representativeness*."

Sample Bias and Sampling

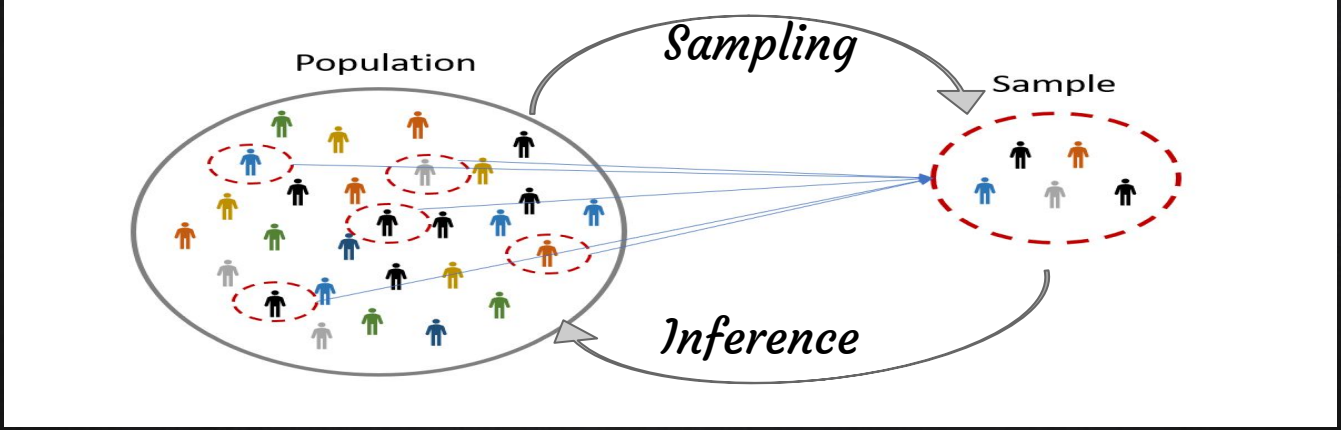
Type of sample bias:

- Selection bias
- Data collection bias
- Prescreen bias
- Non-probability sample bias – unquantifiable

Sampling Designs help overcome bias:

- Random sampling: "Drawing elements into a sample at random"
- Stratified sampling: "Dividing the population into strata and randomly sampling from each strata"
- Simple Random Sample: "The sample was results from random sampling without stratifying"
- Clustered Sample: "occurs when the natural sampling unit is a group or cluster of individual units"
- Systematic Sample: "the selection of every kth element from a sampling frame or from a sequential stream of potential respondents"

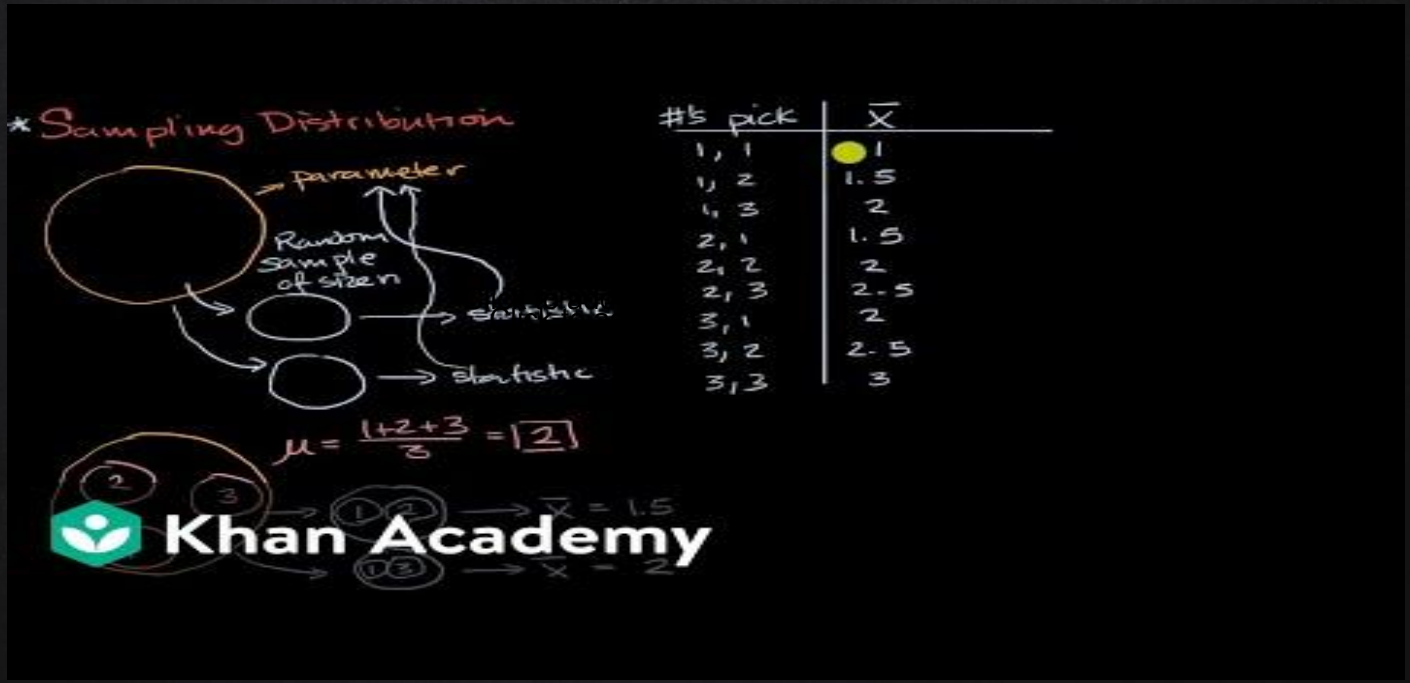
Sample Statistics are Random Variables with Distribution



Population parameters: population mean, population variance

Sample statistics: sample mean, sample variance

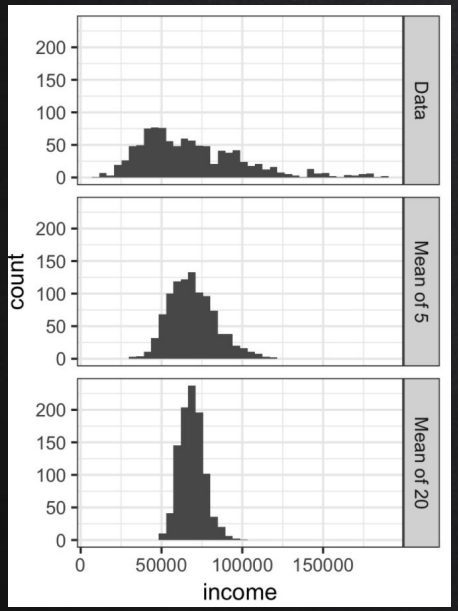
Sampling Distribution



Sampling distribution video from Khan Academy

https://www.youtube.com/watch?v=z0Ry_3_qhDw

Data Distribution and Sampling Distribution



Sample statistic
"A metric calculated for a sample of data drawn from a larger population"

Data distribution
"The frequency distribution of individual *values* in a data set"

Sampling distribution
"The frequency distribution of a *sample statistic* over many samples or resamples"

Central limit theorem
"The tendency of the sampling distribution to take on a normal shape as sample size rises"

Standard Error of the Mean

Mean and standard error of sample mean video from Khan Academy

https://www.youtube.com/watch?v=J1twbrHel3o

Resampling Methods

Bootstrap: Random sample with replacement

Data: A, B, C, D

Samples: (A/B), (B/C), (A/B), (C/D).....

Permutation: Random sample without replacement

Data: A, B, C, D

Samples: (A/B), (A/C), (A/D), (B/C), (B/D), (C/D)

Bootstrap is the most popular resampling method (an non-parametric approach) for parameter estimate as well as some non-parametric ML algorithms

Warning: The bootstrap does not compensate for a small sample size; it does not create new data, nor does it fill in holes in an existing data set. It merely informs us about how lots of additional samples would behave when drawn from a population like our original sample

Table of Contents for Today's Discussion

1. Basics Statistics to understand data (EDA statistics)

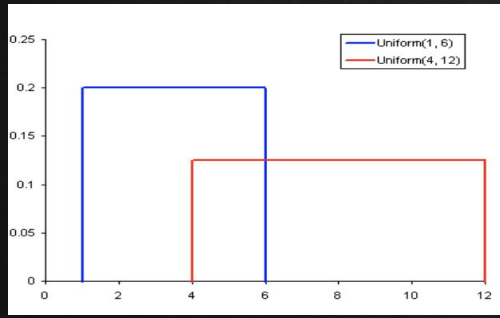
2. Sampling Techniques

3. Probability distribution and Inferential statistical analysis

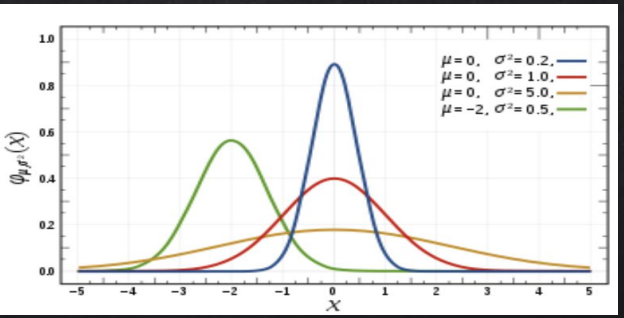
Probability Density Function

Probability is the measure of how likely an event will occur. More precise, probability is the ratio of desired outcomes to total outcomes: *(desired outcomes) / (total outcomes)*

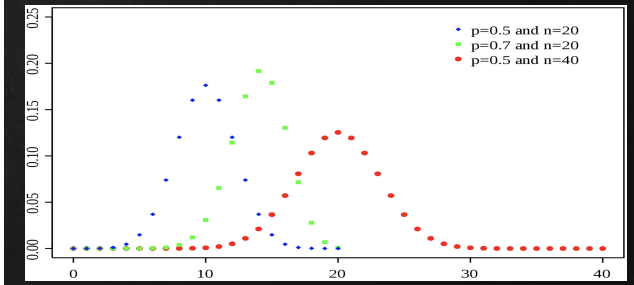
Uniform



Normal



Binomial



Experiment and Inferential Statistical Analysis

A/B Testing: "An A/B test is an experiment with two groups to establish which of two treatments, products, procedures, or the like is superior. Often one of the two treatments is the standard existing treatment, or no treatment. If a standard (or no) treatment is used, it is called the control. A typical hypothesis is that a new treatment is better than the control"

Examples are common in web design and marketing, drug effectiveness

- Testing two soil treatments to determine which produces better seed germination
- Testing two therapies to determine which suppresses cancer more effectively
- Testing two prices to determine which yields more net profit
- Testing two email headlines to determine which produces more clicks
- Testing two web ads to determine which generates more conversions

Statistical Testing

***The Null Hypothesis*:** treatment A has the same response as B

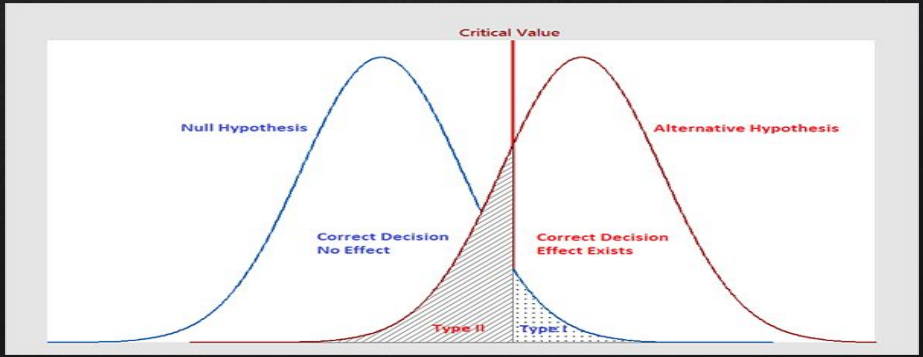
Alternative Hypothesis: treatment A has better response than B

P-value: Under the null hypothesis (Or if the stated hypothesis is true), the p-value is the probability of obtaining a result(statistic) as unusual or extreme as the observed results.

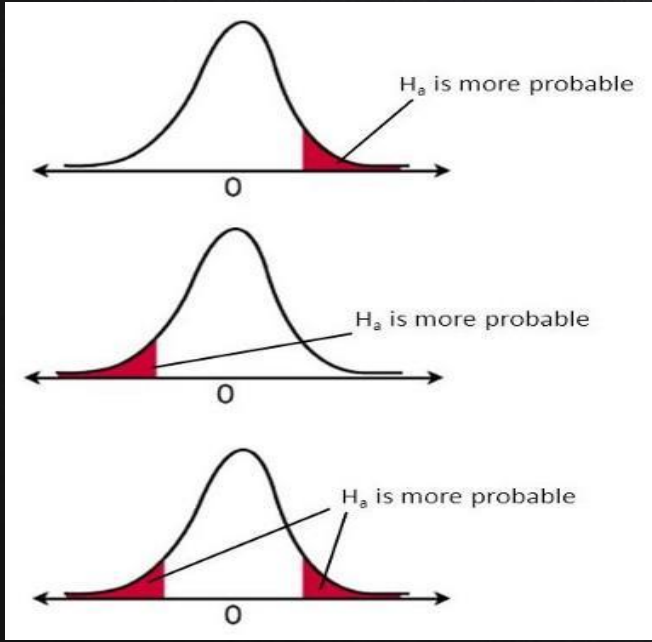
Alpha - significance level: The probability threshold of "unusualness" that chance results must surpass for actual outcomes to be deemed statistically significant. Or Probability of making a Type I Error: Reject null hypothesis when null hypothesis is true

Type 1 error: error occurs when the sample results, lead to the rejection of the null hypothesis when it is in fact true. Type-I errors are equivalent to false positives.

Type 2 error: error occurs when based on the sample results, the null hypothesis is not rejected when it is in fact false. Type-II errors are equivalent to false negatives.



One-Tailed and Two-Tailed Tests



Right-tail test

$H_a: \mu > \text{value}$

Left-tail test

$H_a: \mu < \text{value}$

One Tailed Test

Two-tail test

$H_a: \mu \neq \text{value}$

Two Tailed Test

Test Statistic:

"The **test statistic** measures how close the sample has come to the null hypothesis. Its observed value changes randomly from one random sample to a different sample. A test statistic contains information about the data that is relevant for deciding whether to reject the null hypothesis or not"

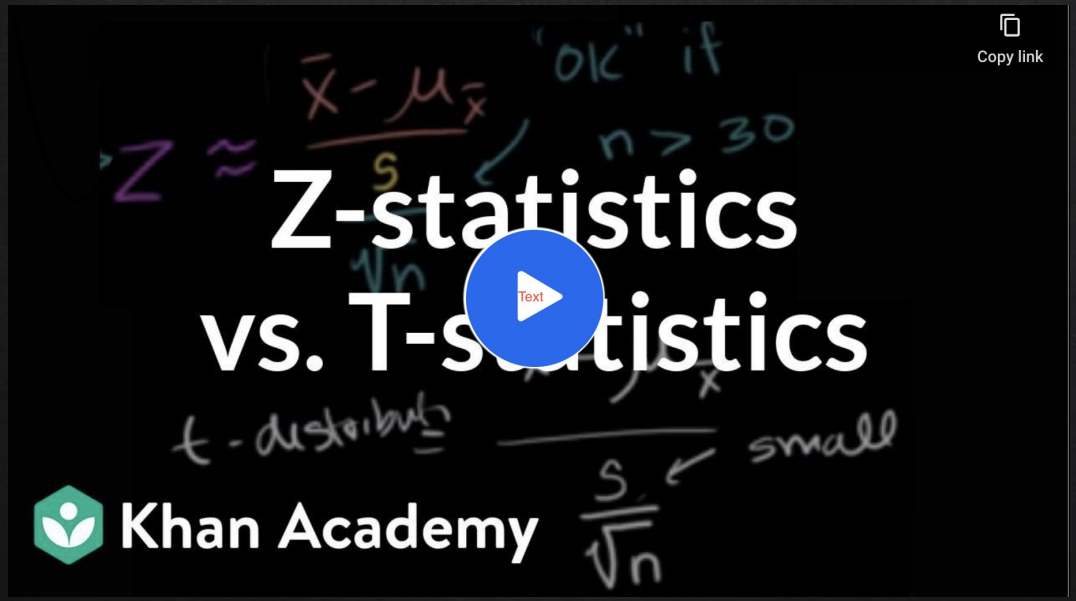
One sample test and test statistic

Hypothesis Test	Test Statistics	Expression
Z-Test	Z-statistics	$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$
T-test	t-statistics	$\frac{\bar{X}-\mu}{S/\sqrt{n}}$

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$$

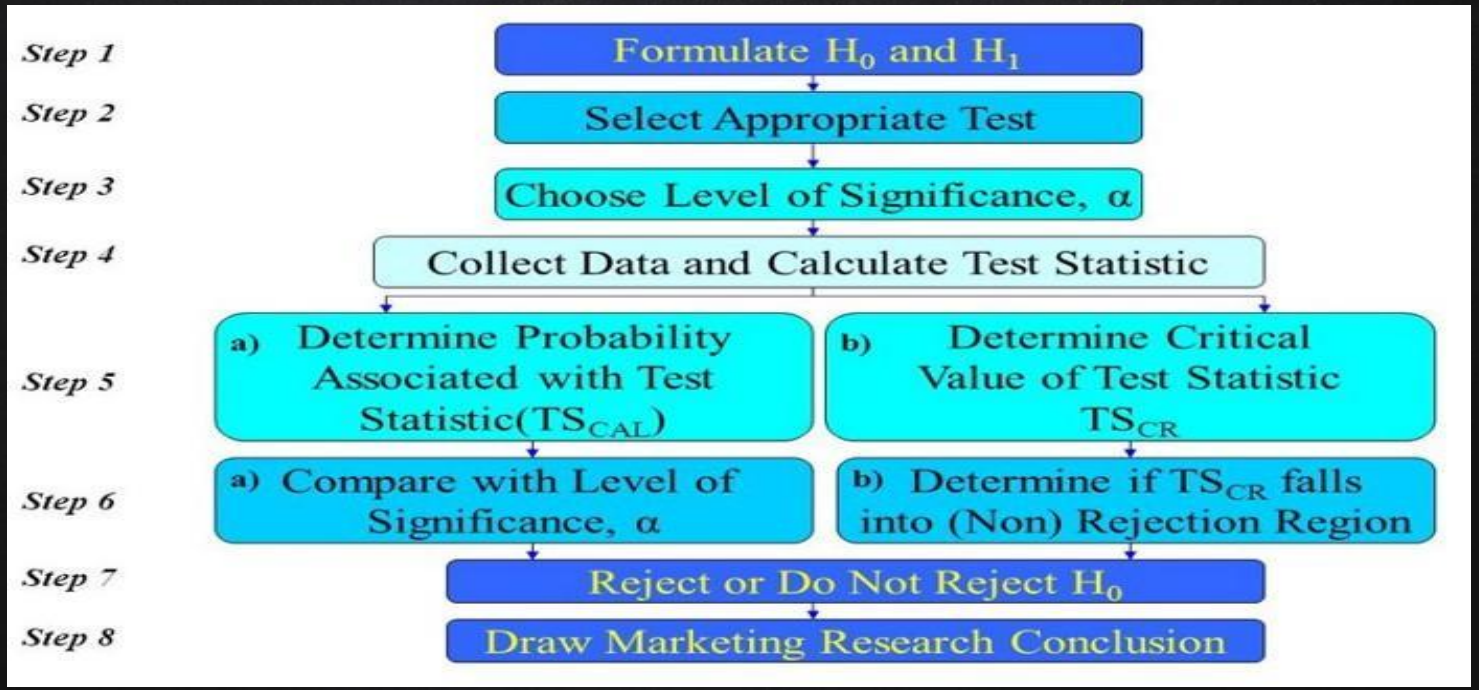
$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

The Difference Between Z-statistics and T-statistics



T and Z statistics video from Khan Academy
Difference of T and Z statistics

The Normal Procedure for AB Testing



Exercises

A major department store is considering the introduction of an Internet shopping service. The new service will be introduced if more than 40 percent of the Internet users shop via the Internet.

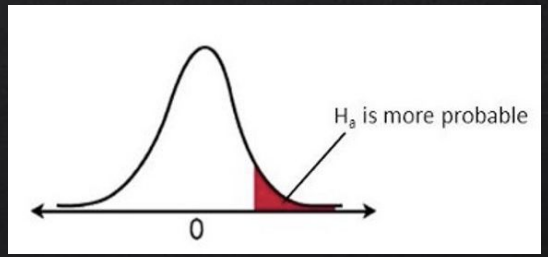
The experiment result: 30 users were surveyed and 17 indicated that they used the Internet for shopping.

Please following the procedure to conduct a hypothesis testing to suggest – should the department store introduce the internet shopping service?

Solution – Step1 Formulating the Hypothesis

$$H_0: \pi \leq 0.40$$

$$H_1: \pi > 0.40$$



Solution – Step2 Select a Suitable Test

This is a proportion problem, binomial distribution and follow the central limit theorem, when n is large enough, the sample mean following normal distribution, then Z test and z-statistic will be the one to use

$$Z = (p-\pi) / \acute{o} \text{ where } \acute{o} = \text{sqrt}(\pi(1-\pi)/n)$$

Solution – Step3 Choose Level of Significance, α

Level of Significance refers to ***Type-I*** error. In this problem, a Type-I error would occur if we concluded, based on the sample data, that the proportion of customers preferring the new service plan was greater than 0.40, when in fact it was less than or equal to 0.40.

On the other hand, ***Type-II*** error should occur when the sample data help concluded that the proportion of customer would prefer the new service is less than or equal to 0.4, while the true proportion is larger than 0.4.

The choice should balance the two types of error, normally 0.05 and 0.01 are commonly used significance levels, other choice are rare

Choice **α = 0.05**

Solution – Step4: Collect Data and Calculate Test Statistics

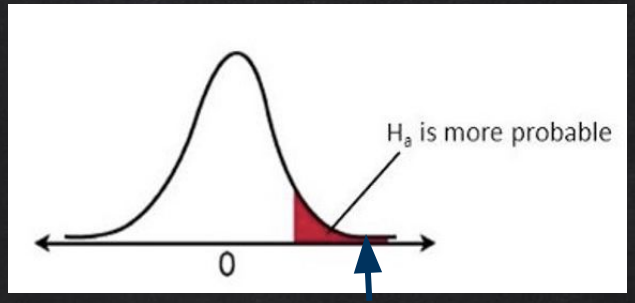
N=30, Yes: 17

Therefore p = 17/30 = 0.567

For ó = sqrt(0.4*(1-0.4)/30) = 0.089

Finally: z = (p-π) / ó
= (0.567 - 0.4) / 0.089
= 1.88

Solution – Step5 Determine the Test Statistic Probability(p-value)



Critical Value

Approach 1:
Pr(z<=1.88) = 0.9699
P-value = 1- Pr(z<=1.88) = 0.03

Approach 2:
Critical Value(CV) (z when Pr(z<=CV)=0.95)
CV = 1.65

Solution – Step 6&7 Compare with Significant Level and Draw Conclusion

P-value = 0.03 < 0.05 = α OR CV = 1.65 < 1.88

Reject the Null hypothesis, accept the alternative hypothesis, the customer proportion of prefer the new service is greater than 0.4

Solution – Step 8 Conclusion and More

This test concluded that there is evidence that the proportion of Internet users who shop via the Internet is significantly greater than 0.40. Hence, the recommendation to the department store would be to introduce the new Internet shopping service.

Assumption of One-Sample Parametric Statistical Testing

- One-Sample Z-Test Assumptions:

  1. The data are continuous (not discrete).

  2. The data follow the normal probability distribution.

  3. The sample is a simple random sample from its population. Each individual in the population has an equal probability of being selected in the sample.

  4. The population standard deviation is known.

  5. Binary data proportion test validated by Central Limit Theorem for normality, with $n*\pi > 5$ and $n*(1-\pi) > 5$

- One-Sample T-Test Assumptions The assumptions of the one-sample t-test are:

  1. The data are continuous (not discrete).

  2. The data follow the normal probability distribution.

  3. The sample is a simple random sample from its population. Each individual in the population has an equal probability of being selected in the sample.

Types of Statistical Testing I

- One sample test

T-Test and Z-test

- Two sample test with equal variance

T-Test and Z-test

- Two sample test with unequal variance

T-Test and Z-test

Detail information

Hypothesis Test	Test Statistics	Expression
One sample Z-Test	Z-statistics	$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$
One sample T-test	T-statistics	$\frac{\bar{X}-\mu}{S/\sqrt{n}}$
Two sample Z-Test equal variance	Z-statistics	$z=\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$
Two sample Z-Test unequal variance	Z-statistics	$z=\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$
Two sample T-Test equal variance	T-statistics	$T=\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}}$
Two sample T-Test unequal variance	T-statistics	$T=\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{S_X^2}{n}+\frac{S_Y^2}{m}}}$

delta1=delta2

Types of Statistical Testing II

Parametric: "assume underlying statistical distributions in the data. Therefore, several conditions of validity must be met so that the result of a parametric test is reliable. For example, Student's t-test for two independent samples is reliable only if each sample follows a normal distribution and if sample variances are homogeneous. "

Nonparametric: "do not rely on any distribution. They can thus be applied even if parametric conditions of validity are not met."

Pros and cons:

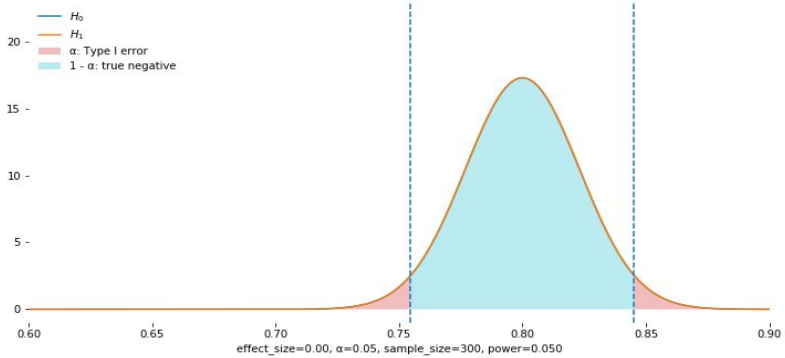
- "Nonparametric tests are more **robust** than parametric tests. In other words, they are valid in a broader range of situations (fewer conditions of validity)."
- "The advantage of using a parametric test instead of a nonparametric equivalent is that the former will have more statistical **power** than the latter. In other words, a parametric test is more able to lead to a rejection of H0. Most of the time, the p-value associated to a parametric test will be lower than the p-value associated to a nonparametric equivalent that is run on the same data."

Power and Sample Size

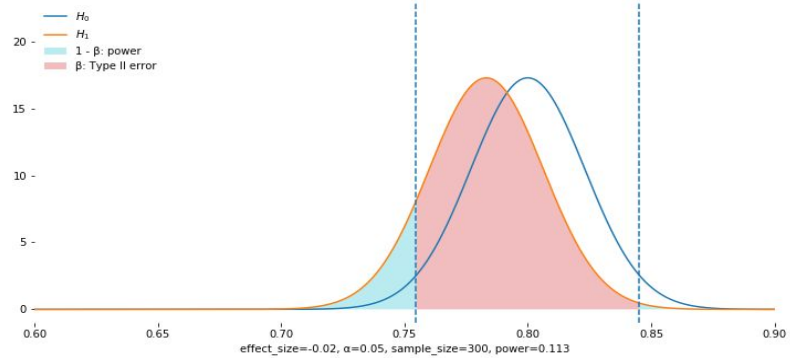
Before we begin the data collection, we should have put several things in consideration forehand to make the most out of a experimental design to fullfill the purpose of our study to prevent:

1. Test was not sensitive because the sample size was not large enough
2. Is the test provide enough power to test the difference if the experiment is truly different

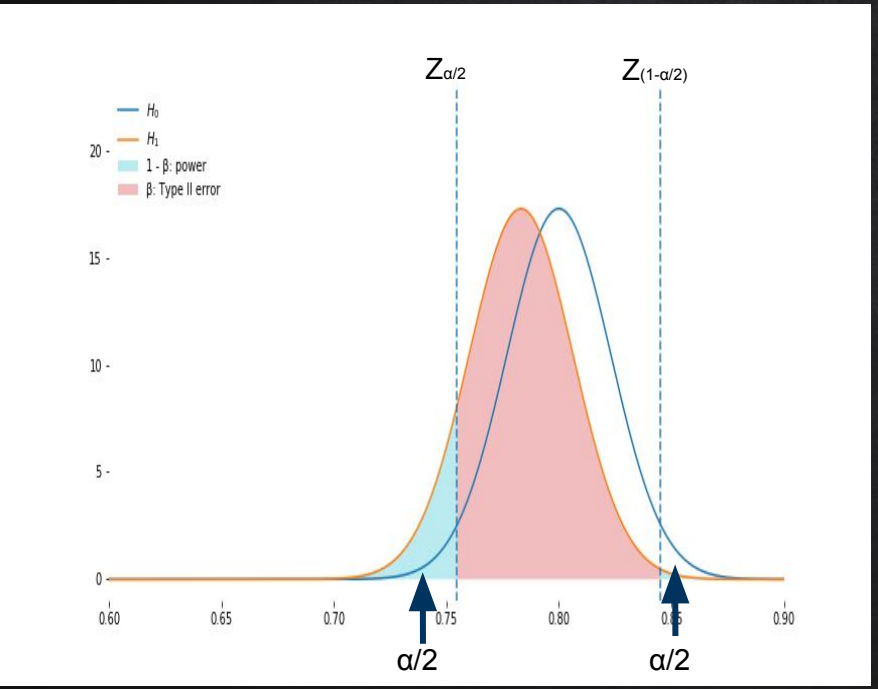
When Null Hypothesis is True



When Null Hypothesis is False



How To Calculate Power



$H_0: \mu = \mu_0 \qquad H_a: \mu \neq \mu_0$

Test Statistic: $Z = \dfrac{(\bar{y}-\mu)}{\sigma/\sqrt{n}} = \dfrac{\sqrt{n}(\bar{y}-\mu)}{\sigma}$

significance level α:

$$\alpha = Pr\{\text{Reject } H_0 / H_0 \text{ True}\} = Pr\{Z < Z_{\frac{\alpha}{2}} \text{ or } Z > Z_{(1-\frac{\alpha}{2})} / \mu = \mu_0\}$$

Power:

$$1-\beta = Pr\{\text{Reject } H_0 / H_0 \text{ False}\} = Pr\{Z < Z_{\frac{\alpha}{2}} \text{ or } Z > Z_{(1-\frac{\alpha}{2})} / \mu \neq \mu_0\}$$
$$= Pr\{Z < Z_{\frac{\alpha}{2}} / \mu \neq \mu_0\} + Pr\{Z > Z_{(1-\frac{\alpha}{2})} / \mu \neq \mu_0\}$$

Assume: H_0 with mean μ_0

H_a with mean μ_a.

$$\text{Power} = 1-\beta = Pr\left\{\frac{\sqrt{n}(\bar{y}-\mu_0)}{\sigma} < Z_{\frac{\alpha}{2}} / \mu = \mu_a\right\} + Pr\left\{\frac{\sqrt{n}(\bar{y}-\mu_0)}{\sigma} > Z_{(1-\frac{\alpha}{2})} / \mu = \mu_0\right\}$$
$$= Pr\left\{\frac{\sqrt{n}(\bar{y}-\mu_a+\mu_a-\mu_0)}{\sigma} < Z_{\frac{\alpha}{2}} / \mu = \mu_a\right\} + Pr\left\{\frac{\sqrt{n}(\bar{y}-\mu_a+\mu_a-\mu_0)}{\sigma} > Z_{(1-\frac{\alpha}{2})} / \mu = \mu_a\right\}$$
$$= Pr\left\{\frac{\sqrt{n}(\bar{y}-\mu_a)}{\sigma} < Z_{\frac{\alpha}{2}} - \frac{\sqrt{n}(\mu_a-\mu_0)}{\sigma} / \mu = \mu_a\right\} + Pr\left\{\frac{\sqrt{n}(\bar{y}-\mu_a)}{\sigma} > Z_{(1-\frac{\alpha}{2})} - \frac{\sqrt{n}(\mu_a-\mu_0)}{\sigma} / \mu = \mu_a\right\}$$

What's The Constraints of Size and Power Calculation

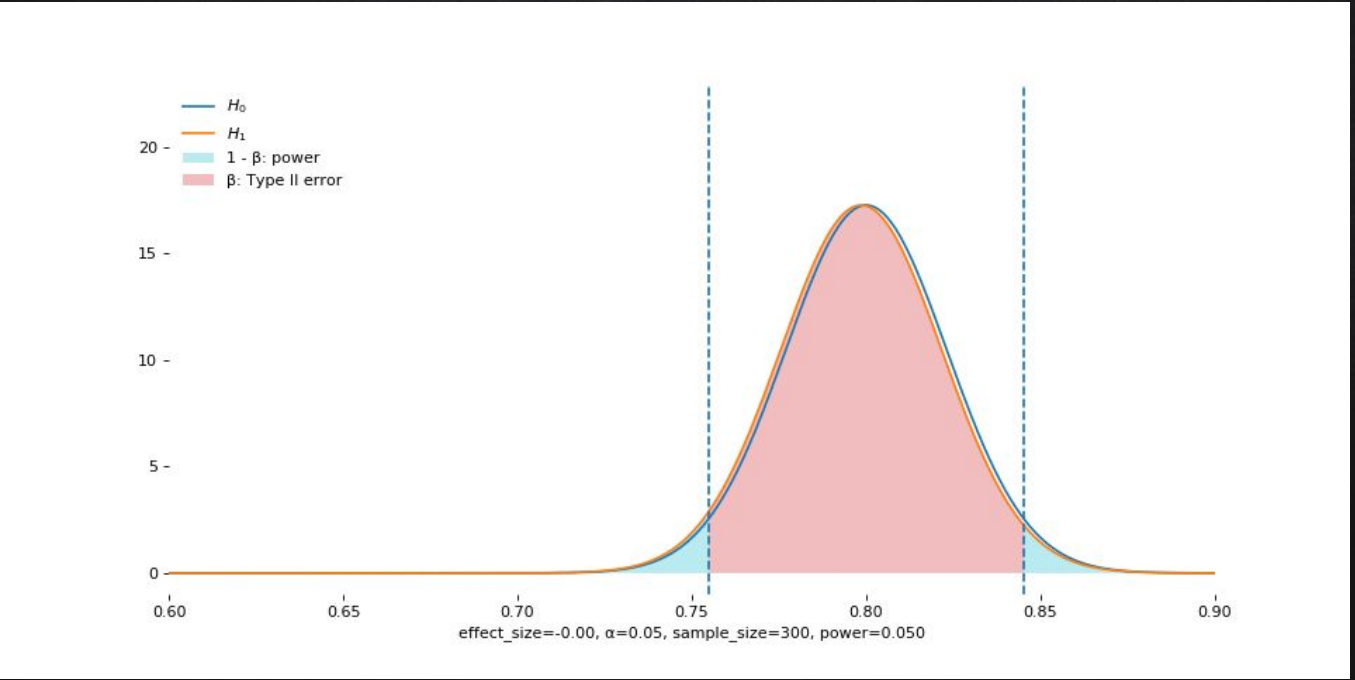
To calculate power or required sample size, there are four moving parts:

- Sample size
- Effect size you want to detect
- Significance level (alpha) at which the test will be conducted
- Power

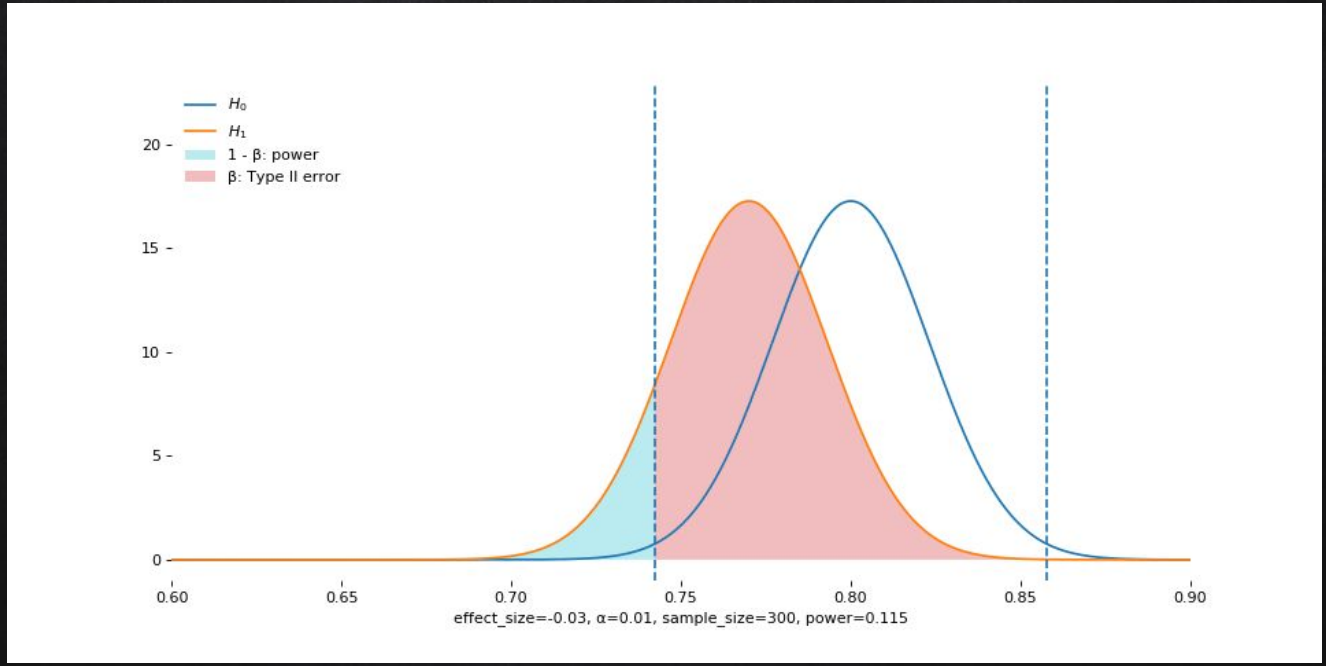
Specify any three of them, and the fourth can be calculated. Most commonly, you would want to calculate sample size, so you must specify the other three.

$$z = (p - p_0)/\text{sqrt}((p_0(1-p_0)/n))$$

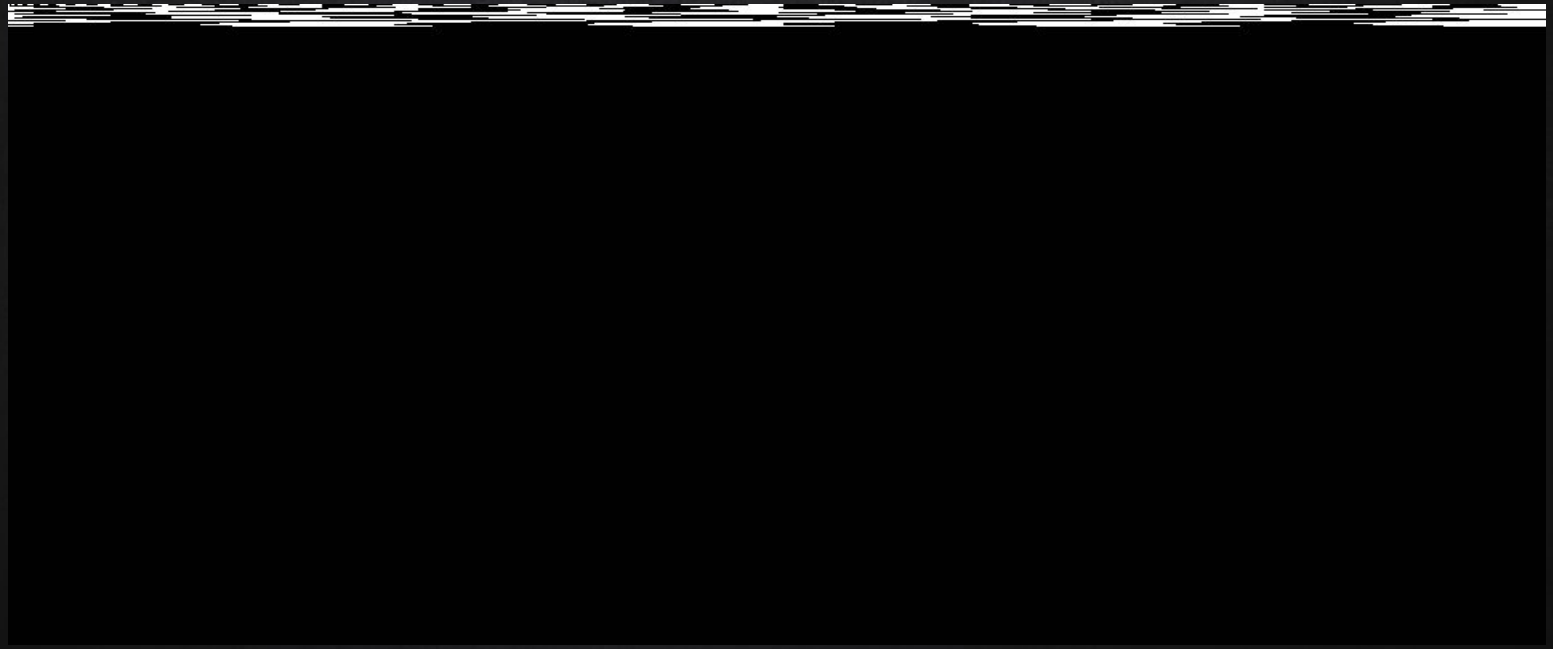
How Effective Size Impact Power



How Alpha Impact Power



How Sample Size Impact Power



General Process for Data Science Project

References

https://thefactfactor.com/facts/management/statistics/variable/1432/

https://towardsdatascience.com/8-fundamental-statistical-concepts-for-data-science-9b4e8a0c6f1c

https://learning.oreilly.com/library/view/Practical+Statistics+for+Data+Scientists,+2nd+Edition/9781492072935/ch03.html#Significance

https://en.wikipedia.org/wiki/Sampling_bias

https://towardsdatascience.com/everything-you-need-to-know-about-hypothesis-testing-part-i-4de9abebbc8a

https://towardsdatascience.com/the-power-of-a-b-testing-3387c04a14e3

https://www.edureka.co/blog/statistics-and-probability/#Inferential%20Statistics

https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779

https://towardsdatascience.com/the-5-basic-statistics-concepts-data-scientists-need-to-know-2c96740377ae